

# Propositions de la TEI pour l'édition numérique des sources primaires

## L'édition numérique : en pratique

On fait une édition numérique à partir des composants suivants :

- un ensemble d'images numériques, chacun représentant une page d'une ressource
- une transcription plus ou moins complète des textes qui figurent sur ces pages
- des métadonnées concernant les ressources numérisées, par exemple leur relation
- des métadonnées concernant la manière de sa numérisation
- des annotations plus ou moins riches sur les entités référencées dans les textes, la langue du texte, etc

Le système TEI nous propose une manière de structurer tout cela

## Structures TEI pour la transcription des sources

- **<text>** : rassemble une lecture structurée du contenu intellectuel d'un document (ou en ensemble des documents) son 'texte'
- **<facsimile>** : regroupe un ensemble d'images représentant les pages (*vel sim*) d'un document
- **<sourceDoc>** : facsimile et transcription quasiment objective des aspects physique d'un seul document
- **<teiHeader>** : fournit des métadonnées qui décrivent les objects traités et les méthodes d'encodage concernées

Un élément **<TEI>** contient au minimum un **<teiHeader>**, suivi d'un ou plusieurs des autres possibilités



## Deux methodes simplissimes pour gerer un "facsimilé numérique"

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- metadonnees sur l'édition numérique -->
  </teiHeader>
  <facsimile>
    <graphic url="page1r.png"/>
    <graphic url="page1v.png"/>
    <graphic url="page2r.png"/>
    <graphic url="page2v.png"/>
  </facsimile>
</TEI>
```

NB: aucune maniere de structurer les images

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- metadonnees sur l'édition numérique -->
  </teiHeader>
  <text>
    <pb facs="page1.png"/>
    <!-- texte de la page 1 facultativement transcrit ici -->
    <pb facs="page2.png"/>
    <!-- texte de la page 2 facultativement transcrit ici-->
  </text>
```

## Inconvénients

- difficile de supporter des relations plus compliquées
- maintien difficile d'informations propres à l'image
- nécessite donc l'intégration avec d'autres fichiers (typiquement METS)
- plusieurs images pour un seul page ?
- besoin d'associer des pages, par ex pour une feuille, ou une double-page

## On dispose de versions alternatives d'une même image

L'élément `<surface>` nous permet de regrouper les images équivalentes:

```
<facsimile>
  <graphic url="page1.png"/>
  <surface>
    <graphic url="page2-highRes.png"/>
    <graphic url="page2-lowRes.png"/>
  </surface>
  <graphic url="page3.png"/>
  <graphic url="page4.png"/>
</facsimile>
```

## On veut identifier les feuilles d'un manuscrit

L'element <surfaceGrp> nous permet de regrouper les surfaces:

```
<facsimile>
  <surfaceGrp type="leaf">
    <surface>
      <graphic url="pagelrecto.png"/>
    </surface>
    <surface>
      <graphic url="pagelverso.png"/>
    </surface>
  </surfaceGrp>
</facsimile>
```

## On veut distinguer des sous-parties d'une surface

L'element <zone> nous permet d'identifier n'importe quelle region d'une surface.

- Un <zone> identifie un polygone (pas forcement rectangulaire) : une espace en 2d
- Il est defini ou bien par l'attribut @points ou bien par les attributs @ulx, @uly, @lrx et @lry
- Toute definition de zone doit utiliser le *système de coordonnées* definie pour la surface
- Une système de coordonnées definit une plage de valeurs pour les coordonnees (x,y) qui expriment un polygon en 2d
- (il ne s'agit pas d'une mensuration)

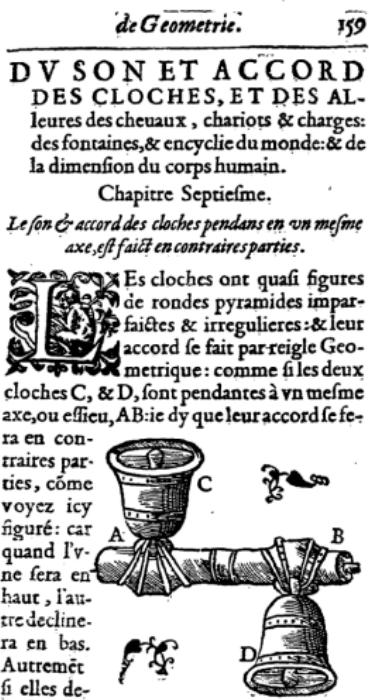
```
<facsimile>
  <surface ulx="0" uly="0" lrx="40" lry="30">
    <graphic url="page1r.png"/>
    <zone points="22,10 30,21 17,25 12,23">
      <graphic url="page1rdetail.png"/>
    </zone>
  </surface>
</facsimile>
```

## Lier les images avec leur transcription

- L'attribut *@facs* est disponible sur tout élément de transcription. Il pointe sur un `<zone>`, `<surface>`, ou (plus simplement) sur un `<graphic>`
- (L'attribut *@start* de `<zone>` ou de `<surface>` pointe dans l'autre sens, vers une portion de transcription)

```
<facsimile>
  <surfaceGrp type="leaf">
    <surface xml:id="plr">
      <graphic url="page1r.png"/>
      <graphic url="page1r.tiff"/>
    </surface>
    <surface xml:id="plv">
      <graphic url="page1v.png"/>
    </surface>
  </surfaceGrp>
</facsimile>
<text>
  <pb facs="#plr"/>
  <!-- text from page 1 recto transcribed here -->
  <pb facs="#plv"/>
  <!-- text from page 1 verso transcribed here -->
</text>
```

# Exemple



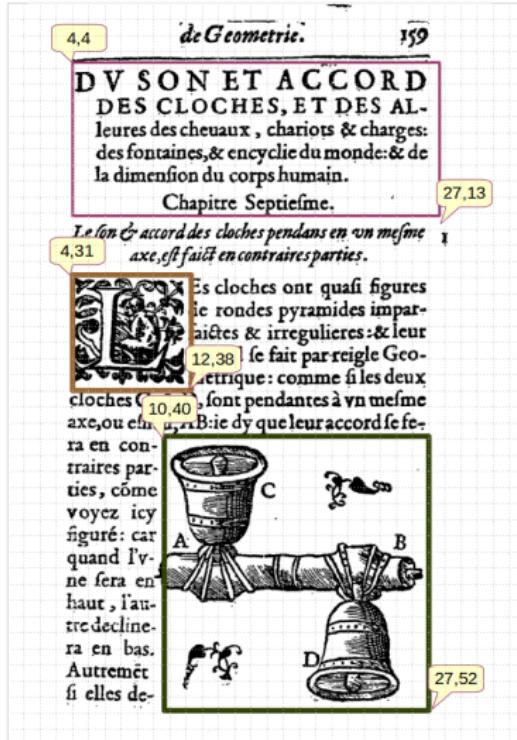
Nous distinguons plusieurs zones  
sur cette surface :

- le titre
- la lettrine
- l'image d'une cloche

...

Ce sont tous des zones  
rectangulaires

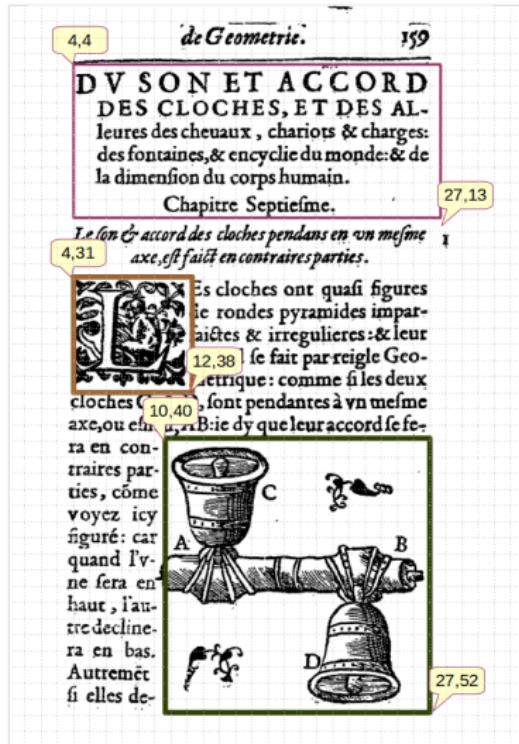
# Exemple



```
<facsimile>
  <surface xml:id="B49r" ulx="0" uly="0" lrx="52"
    <graphic url="bovelles.png"/>
    <zone xml:id="B49rHead" ulx="4" uly="4" lrx="10"/>
      <!-- le titre -->
    <zone xml:id="B49rCap" ulx="4" uly="31" lrx="10"/>
      <!-- la lettrine -->
    <zone xml:id="B49rFig" ulx="10" uly="40" lrx="52"/>
      <!-- la cloche -->
    </surface>
  </facsimile>
```

(Notons que nous introduisons  
des identifiants pour chaque  
zone)

# Et la transcription

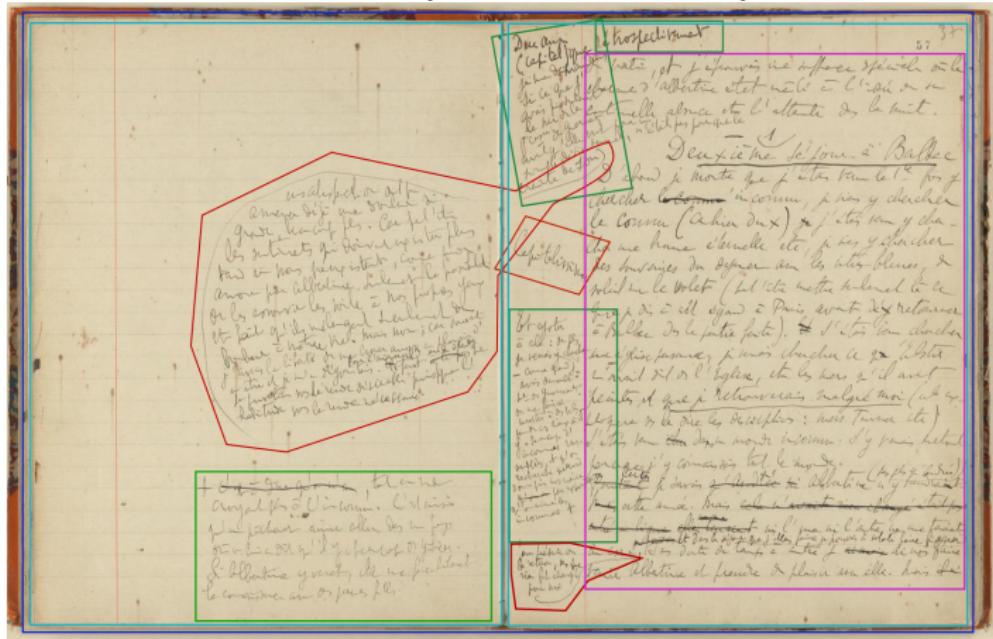


```
<facsimile>
  <surface xml:id="B49r" ulx="0" uly="0" lrx="52"
    <graphic url="bovelles.png"/>
    <zone xml:id="B49rHead" ulx="4" uly="4" lrx="10"/>
    <!-- le titre -->
    <zone xml:id="B49rCap" ulx="4" uly="31" lrx="10"/>
    <!-- la lettrine -->
    <zone xml:id="B49rFig" ulx="10" uly="40" lrx="10"/>
    <!-- la cloche -->
  </surface>
</facsimile>
<text>
  <body>
    <pb facs="#B49r"/>
    <fw>De Geometrie 159</fw>
    <head facs="#B49rHead"> DU SON ET ACCORD  
DES CLOCHES ET DES ALLEURES DES CHEUAUX,  
CHARIOTS & CHARGES: DES FONTAINES,&  
ENCYCLIE DU MONDE: & DE LA  
DIMENSION DU CORPS HUMAIN.</head>
    <head facs="#B49rCap">Chapitre Septiesme.</head>
    <div n="1">
      <p>Le son & accord des cloches pendans en  
ung mesme axe, est faict en  
contraires parties.</p>
      <p><g facs="#B49rCap">L</g>Es cloches ont  
quasi figures de rondes pyramides  
imperfaites & irregulieres: &  
leur accord se fait par reigle  
Geometrique: comme si les deux  
cloches C & D sont pendantes à vn mesme axe,  
ou essieu, A B: je dy que leur  
accord se fera en contraires parties  
co<ex>m</ex>me voyez icy figuré.  
Car quand l'vene sera en haut, l'autre declinera  
en bas. Autremēt<ex>m</ex>t si

```

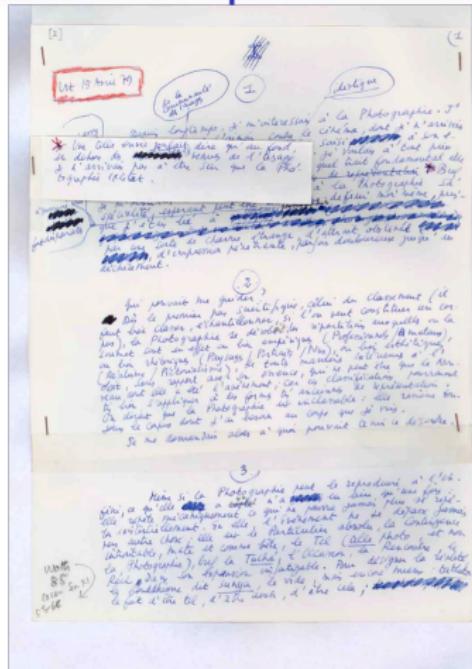
# Surfaces et zones...

La relation surface/zone peut être très complexe :



Source gallica.bnf.fr / Bibliothèque nationale de France

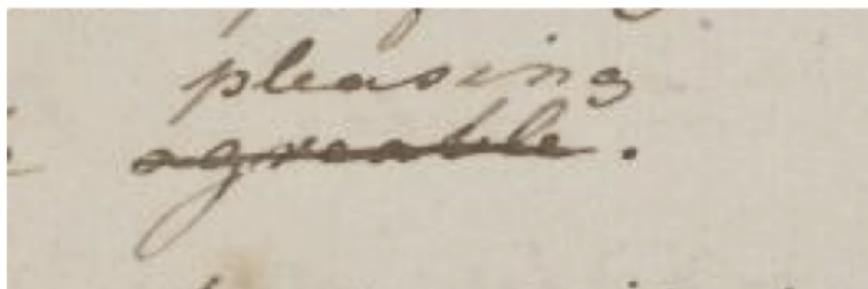
... les zones peuvent traverser les frontières des surfaces



Elena va discuter de cela plus tard!

## Transcription : une boite de Pandore

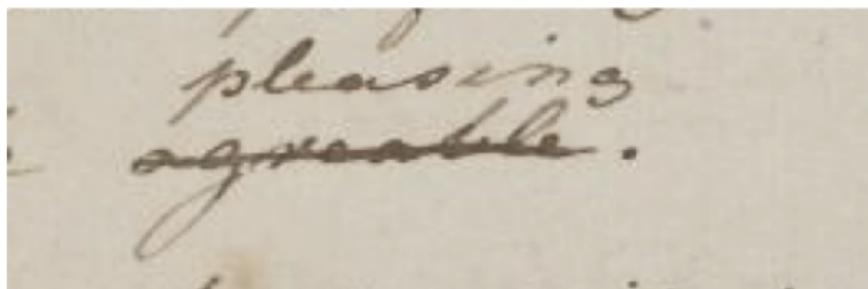
Qu'est-ce qui se passe ici?



- ➊ 'agreeable' est biffé 'pleasing' est écrit au dessus dans l'espace interlineaire.
- ➋ 'agreeable' est rature, et remplace par 'pleasing'
- ➌ Initialement, le texte lisait 'agreeable', mais ce mot a été supprimé, et à un temps ultérieur le mot 'pleasing' a été ajouté.

## Transcription : une boite de Pandore

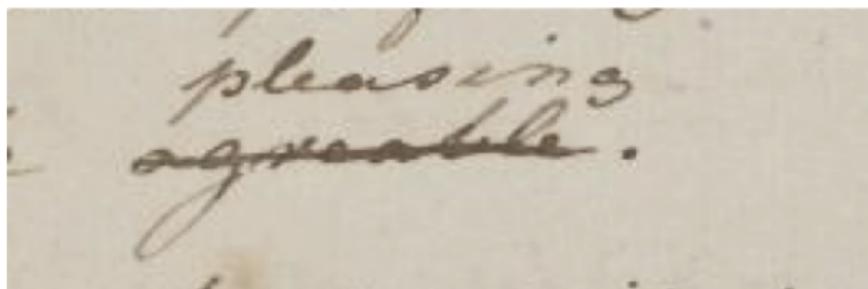
Qu'est-ce qui se passe ici?



- ➊ 'agreeable' est biffé 'pleasing' est écrit au dessus dans l'espace interlineaire.
- ➋ 'agreeable' est rature, et remplace par 'pleasing'
- ➌ Initialement, le texte lisait 'agreeable', mais ce mot a été supprimé, et à un temps ultérieur le mot 'pleasing' a été ajouté.

## Transcription : une boite de Pandore

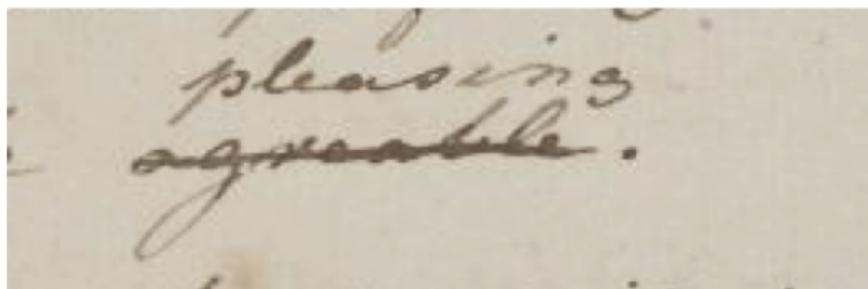
Qu'est-ce qui se passe ici?



- ① 'agreeable' est biffé 'pleasing' est écrit au dessus dans l'espace interlineaire.
- ② 'agreeable' est rature, et remplace par 'pleasing'
- ③ Initialement, le texte lisait 'agreeable', mais ce mot a été supprimé, et à un temps ultérieur le mot 'pleasing' a été ajouté.

## Transcription : une boite de Pandore

Qu'est-ce qui se passe ici?



- ① 'agreeable' est biffé 'pleasing' est écrit au dessus dans l'espace interlineaire.
- ② 'agreeable' est rature, et remplace par 'pleasing'
- ③ Initialement, le texte lisait 'agreeable', mais ce mot a été supprimé, et à un temps ultérieur le mot 'pleasing' a été ajouté.

## Transcription: une espece de lecture

Quels sont les buts de la transcription?

- rendre accessible une ressource primaire ...
- ... et comprehensible
- ce qui peut bien impliquer l'addition de beaucoup d'information

Donc

- toute transcription nécessite une sélection parmi les évidences
- toute transcription serait un exercice de l'imaginative

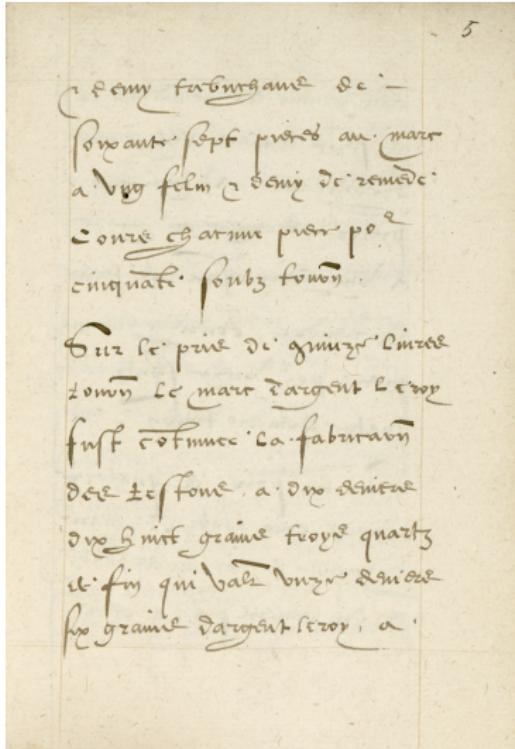
La TEI permet de distinguer la transcription documentaire et la transcription textuelle

## La transcription documentaire

L'element `<sourceDoc>` element permet la representation des lectures 'innocentes' de la texte d'un document

- L'element `<sourceDoc>` peut contenit des `<surface>` et des `<zone>`, pareille a `<facsimile>` ...
- ... sauf que ces composants peuvent contenir de texte transcrive pour accompagner (ou remplacer) les images
- L'element `<line>` (une specialisation de `<zone>`) est disponible
- Et aussi un petit ensemble de balises pas trop semantique, pour noter les interventions metatextuelles (raturage, annotation, etc.)

# Couches de transcription



5

Ensay fabriquame de -  
Prise au sept pices au mar  
a vng film et enay de remede  
contre charme pierre po  
cinqant. pouys fouron.

Sur le pice de quinze linteau  
douz le maro dargent le roay  
fust estoilee la fabriquant  
de la stome a uno de meie  
dne graine foye quartz  
et fin qui valz vingt deniers  
la graine dargent le roay a

- Couche palaeographique : quelles caractères voit-on ?
- Couche documentaire : quels mots (etc) sont identifiables ?
- Couche semantique : comment lire/comprendre cela ?

## Couche palaeographique

- identifier les traces que nous considerons comme des lettres
- faire le mapping entre chacun de ces lettres et un caractere Unicode approprie
- le cas echant, decider quels caracteres non-standard nous souhaitons conserver

L'element `<g>` element est la pour nous aider !

# Une transcription documentaire

5

¶ demy fabrichans de -  
soixante sept pieces au marc  
a ung felin & demy de remede  
Cours chacune piece po&#xFFD;  
cinquâte soubz tourois.  
Sur le pris de quinze livres  
tourois le marc dargent le roy  
fust cõtinuee la fabricaciô  
des testons a dix deniers  
dix huict grains troyz quartz  
et fin qui vaut unze deniers  
ses grains dargent le roy a

```
<surface n="5r">
<zone>5</zone>
<line>& demy trebuchans de</line>
<line>soixante sept pieces au marc</line>
<line>a ung felin & demy de remede</line>
<line>Cours chacune piece po&#xFFD;</line>
<line>cinquâte soubz tourois.</line>
<line>Sur le pris de quinze livres </line>
<line>tourois le marc dargent le roy </line>
<line>fust cõtinuee la fabricaciô</line>
<line>des testons a dix deniers </line>
<line>dix huict grains troyz quartz</line>
<line>de fin qui vaut unze deniers </line>
<line>six grains dargent le roy, a</line>
</surface>
```

## Text-oriented (logical) transcriptional elements

Traditional TEI structuring elements (`<div>`, `<head>`, `<p>` etc.)

Various other phenomena which commonly attract editorial attention :

- original layout information
- abbreviations or other arcana
- 'evident' errors which invite correction or conjecture
- scribal additions, deletions, substitutions, restorations
- non-standard orthography (etc.) which invites normalisation
- irrelevant or non-transcribable material
- passages which are damaged or illegible

## Original layout information

Within `<text>` the logical view is privileged, but the physical view can 'show through' as empty milestone elements :

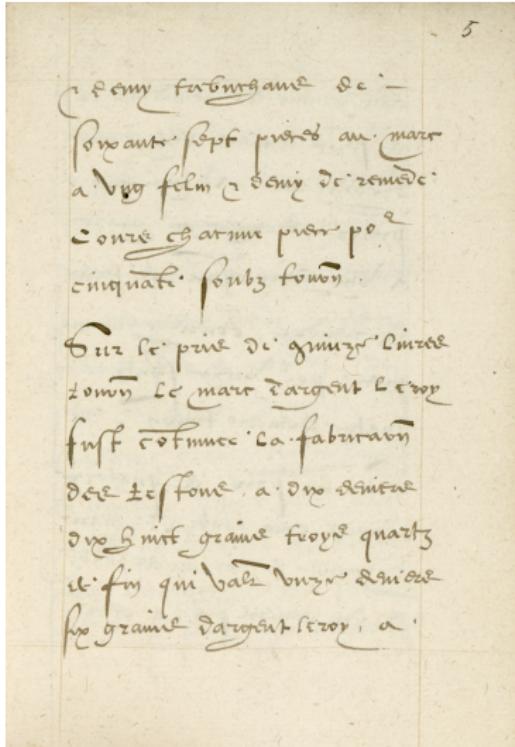
- `<gb>` the start of a new gathering or quire
- `<pb>` the start of a new page
- `<cb>` the start of a new column
- `<lb>` the start of a new written line

These are primarily useful to establish a reference system.

The `<fw>` element can be used to mark 'paratextual' features such as running heads, foliotation etc.

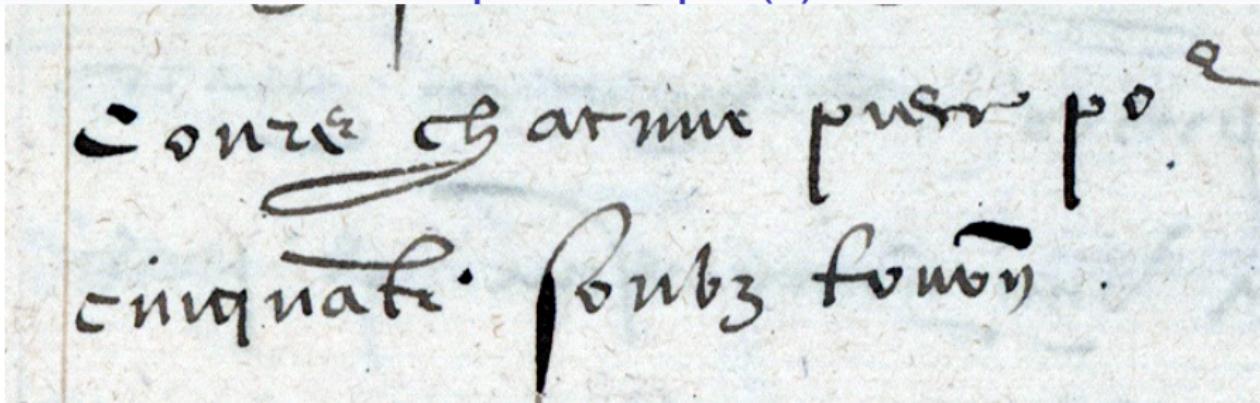
The `<handShift>` element can be used to mark changes of hand or writing in a document.

## Textual transcription of page 5



```
<p>
<!-- ... -->
<pb n="5r"/>
<fw place="topRight" type="pageNum">5</fw>
<lb/>
<expan>et</expan> demy trebuchans de
<lb/>soixante sept pieces au marc <lb/>a ung
felin <expan>et</expan> demy de remede
<lb/>Cours chacune piece <expan>pour</expan>
<lb/>
<expan>cinquante</expan> soubz
<expan>tournois</expan>
<pc>.</pc>
</p>
<p>
<lb/>Sur le pris de quinze livres <lb/>
<expan>tournois</expan> le marc dargent le
roy
<lb/>fust <expan>continuee</expan> la
<expan>fabricacion</expan>
<lb/>des testons a dix deniers <lb/>dix
huict grains troys quartz <lb/>de fin qui
<expan>valent</expan> unze deniers <lb/>six
grains dargent le royst, a
<!-- ... -->
</p>
```

## A simple example (1)



Cours chacune piece pour  
cinquante soubz tournois

Editorial strategy may be simply to note that we have expanded the abbreviations:

```
<p>
<lb/>Cours chacune piece <expan>pour</expan>
<lb/>
<expan>cinquante</expan> soubz <expan>tournois</expan>
<pc>.</pc>
</p>
```

## A simple example (2)

As you noticed, 'pour' was actually written 'po' followed by an 'r' subscript; 'cinquante' as 'cinquāte' with a macron on the 'a' to indicate nasalisation.

We could therefore encode as follows:

```
<p>
  <abbr>po&#xFFD ;</abbr> . . .
  <abbr>cinquāte</abbr>
</p>
```

... or we could choose one of the following styles:

```
<p> po<am>&#xFFD ;</am> . . .
    or po<ex>u</ex>r </p>
```

```
<abbr>po<am>&#xFFD ;</am>
</abbr>
```

```
<expan>po<ex>u</ex>r</expan>
```

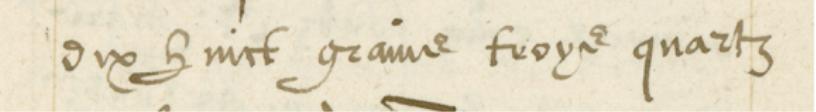
## Simple example (3)

And of course TEI permits both cake and the eating off it:

```
<p> po<choice>
  <am>&#FFFD;</am>
  <ex>ur</ex>
</choice>
</p>
```

```
<choice>
  <abbr>po<am>&#xFFFD;</am>
  </abbr>
  <expan>po<ex>u</ex>r</expan>
</choice>
```

## Normalisation example



dix huit graine troye quartz

```
<lb/>dix <choice>
<orig>huict</orig>
<reg>huit</reg>
</choice> grains
<choice>
<orig>trois
    quartz</orig>
<reg>trois-quart</reg>
</choice>
```

In this case, a further semantic regularisation is possible :

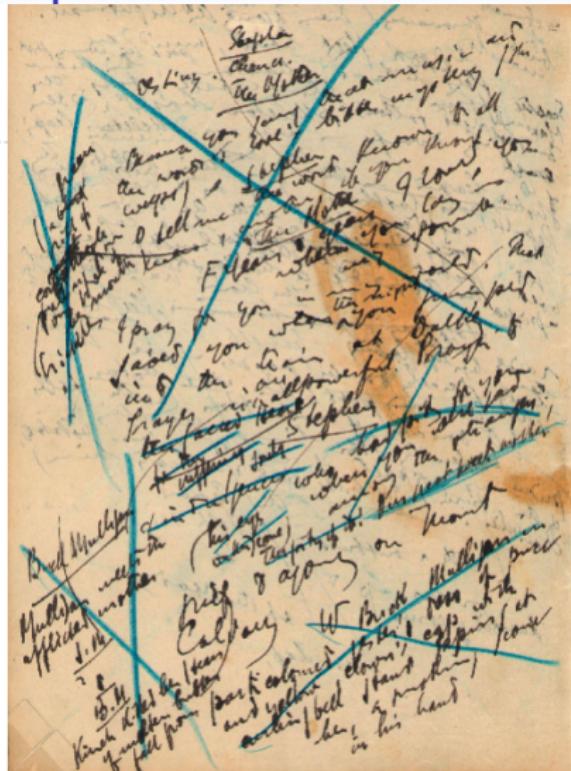
```
<lb/>
<measure quantity="18.75" unit="gr">dix
<choice>
<orig>huict</orig>
<reg>huit</reg>
</choice> grains <choice>
<orig>trois
    quartz</orig>
<reg>trois-quart</reg>
</choice>
<measure>
```

## How far will the TEI take us ?

In particular, is the TEI scheme adequate for the needs of those transcribing 'modern' manuscripts ?

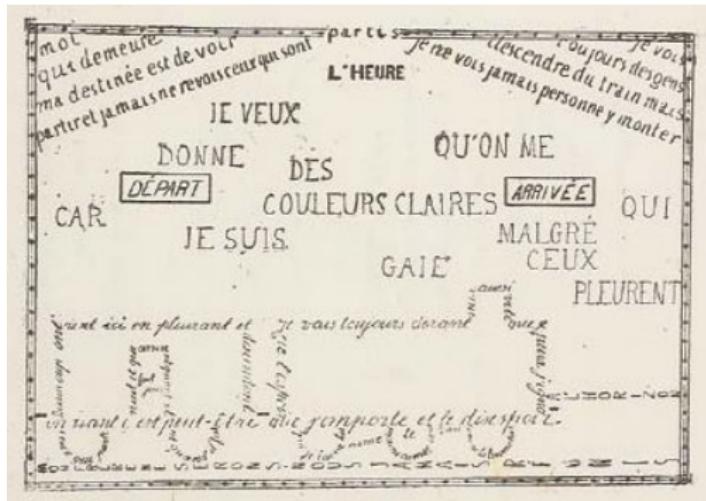
- surviving medieval or early modern manuscripts generally have a public function, and a more or less conventionalised (if complex) format
- modern manuscripts or authorial drafts however often contain entirely private or idiosyncratic signs, with no clear communicative function

For example...



## Text/Image

At all periods we find 'playful' texts whose meaning is conveyed by their documentary appearance as much as by their linguistic properties, or by the interplay between the two.



The TEI initially ruled such texts out of scope, for a variety of reasons, not least a dearth of image-processing technologies.

## Concerns that won't go away

- The process by which a document was created may be as important as its final or canonical textual form
- It may be impossible to talk about the text independently of its documentary instantiation, whether because
  - the meaning of the text is presented entirely or partly graphically
  - the document is deliberately constructed in a non-linear or combinatorial way, in order to generate many 'texts'

## Document vs. text

By hypothesis, distinguishing these levels may help our editorial task :

- at the documentary level : pages, surfaces, writing, tears, crossings-out, stains...
- at the textual level : corrections, modifications, additions, deletions, transpositions...

Distinguishing these levels in our encoding is a good way of studying their interaction